Drought Prediction Using Data Analytics

[1]Fathima G, [2]Chandru M, [3]Chidambaram T, [4]Ariprakash R

[1]Professor, [2]Student, [3]Student, [4]Student
[1]Department of Computer Science and Engineering,
[1]Adhiyamaan College of Engineering, Hosur, India

***Abstract*: Groundwater is the major source of freshwater for drinking, irrigation and industrial purposes. Water Scarcity plays a major problem in rural and urban areas. and the groundwater system's health is reflected in the groundwater levels of the region. There is a need to understand the groundwater scenario with respective regions for the use and management of groundwater resources. Last decades groundwater table level data are analyzed using statistical and arithmetical solutions that are used to predict and manage groundwater level of current and future. The analyzed results can be used to decision making for the water related problems. It will help for overall ground water management.**

***Keywords*: ARIMA (Auto Regression Integration Moving Average), SARIMAX (Seasonal Auto Regression Integration Moving Average)**

I. Introduction

Water is the elixir of life and is essential for sustainable development. Previously it was considered to be an endless natural resource or at least completely renewable. However, in the last 20 years or so, there has been tremendous pressure on this valuable natural resource mainly due to rapid industrialization and population growth. This is because an increase in population will only lead to an increase in the demand for irrigation to meet the needs of food production. Nevertheless, advances in agricultural technology have attracted attention in many regions, illegal irrigation management has resulted in massive deforestation, degraded soils and a deterioration of water quality which makes future water supply unreliable. Keeping in mind the shortage of available water resources in the near future and its future threats, it has become important for water science and planners to reduce their existing water resources. Therefore, a ready reference to monitoring groundwater levels in advance is the need of the hour to carry out continuous water management.

It is required that the prediction is made with great accuracy. The last decade's data in the respective regions is analyzed using mathematical and statistical solutions. The combination of auto regression is a standard sample compilation that will help predict the ground water table level in the coming years. Accuracy can be refined by visualizing the model prediction data of previous years. The drought is mapped using water level fluctuations. This will help predict droughts and therefore the necessary precautions can be taken ahead of time.

II. Related work

Literature survey

Groundwater is major source of our country. India tops the list of using ground water. Ground water is important for sustainability of ecosystem. Groundwater level prediction in India is most importance as our large population is mostly dependent on groundwater for daily consumption. The following comparison is shown below:

The study in [1] used auto-regressive integrated moving average (ARIMA) using Box–Jenkins methodology and fuzzy time-series analysis for forecasting groundwater level. The (ARIMA) model takes care of the seasonal variation of groundwater level in addition to the trend component. The Box–Jenkins technique is used for stationary time series, i.e. it must have a constant mean, variance and autocorrelation.

The study in [2] used autoregressive integrated moving average (ARIMA) model to predict the product's future price. The model was trained with 90% of the collected data and used for predicting the prices of the remaining data. With fine-tuning the ARIMA model parameters (p, q, d) the prediction is performed with a low Mean Absolute Percentage Error Score. ARIMA depends on different inputs such as ACF (Auto Correlation Function) and PACF (Partial Auto Correlation Function) and it works well with the time series theory, if the process is stationary.

The study in [3] used Artificial Bee Colony Algorithm and Back-propagation Neural Network with double hidden layers. A back propagation (BP) neural network based on the artificial bee colony (ABC) optimization algorithm was used to predict groundwater levels in the overexploited arid areas of Northwest China.

The study in [4] used neuro-fuzzy (NF) and artificial neural network (ANN) to predict the groundwater level. Artificial neural networks (ANN) and neuro-fuzzy (NF) models uses time series information of the wells of the respective regions along with its climatic information.

III. Proposed System

Groundwater systems are highly heterogeneous with dynamic temporal-spatial patterns, which makes it difficult to quantify their complex processes, but reliable estimations of groundwater levels are usually required to manage water resources to ensure proper service of water demands in an area. In this study, we developed an analysis that aims to help the ground water body authority which can be used to control and to predict drought earlier. Estimation and usage of ground water resources can be predicted with the help of the last decade water table level data. Opted region for the analysis are Chennai and Dharmapuri. The datasets are collected from the Tamil Nadu water board authority. After pre-processing and analysis of the collected data a predictive model that could deliver approximate ground water availability is established. This would help to a sustainable management of the ground water resource and thus would help in managing drought.

IV. Methodology

Time Series Analysis and ARIMA

Time series is defined as a sequence of time-viewed data or a time series of data points collected from time to time. These are analyzed in order to establish a long-term trend to predict the future or to undertake an alternative analysis. ARIMA models are a set of robust representational models that have the capability to represent both stationary and non-stationary time series. It can produce accurate predictions based on historical data interpretation of single variables. The use of ARIMA for forecasting time series is essential with uncertainty as it does not take knowledge of any underlying model or relationships as in some other forecasting methods. ARIMA essentially depends on past values of the series as well as previous error terms for forecasting. However, ARIMA models are more rigid and more efficient than complex architectural models for short-term estimation. ARIMA stands for Auto Regressive Integrated Moving Average. It is a concept of simpler Auto Regression Moving Average and adds a sense of integration. The acronym is self-explanatory, incorporating important aspects of the model itself. In short, they are:

- AR:(Autoregression) is a model that uses a dependent relationship between one observation and some backward observation.
- I:(Integrated) The variance of raw observations to make the time series consistent (e.g. subtracting from an observation in a previous time step).
- MA: (Moving Average) A model that uses the dependence between observation and residual error from the moving average model applied to lagging observations.

Each of these components is explicitly specified as a parameter in the model. The standard notation of ARIMA (p, d, q) is used, where parameters are substituted with integer values to quickly indicate the use of a particular ARIMA model. The parameters of the ARIMA model are defined as follows:

- p: number of lagged observations considered in the model, also called the lag order.
- d: number of times the observations are differentiated, known as the degree of differentiation.
- q: moving average window size, also called the order of the moving average.

Seasonal ARIMA(SARIMAX)

As the name implies, this model is used when the time series shows the season behavior. This model is one of the classes of the ARIMA model. The only difference is that SARIMAX is added with the additional parameters of the seasonal descriptions.

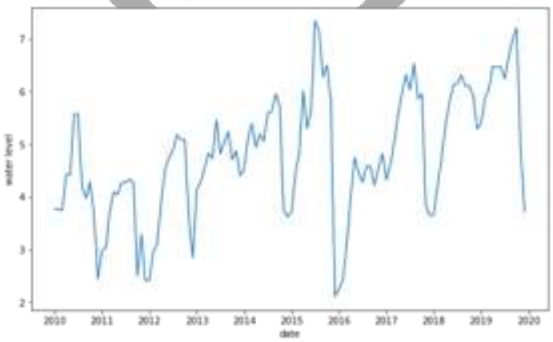
SARIMAX is written as ARIMA (p, d, q) (P, D, Q) m where,

- p: autoregressive value
- d: degree of differencing
- q: number of moving average terms used
- m: number of periods in each season

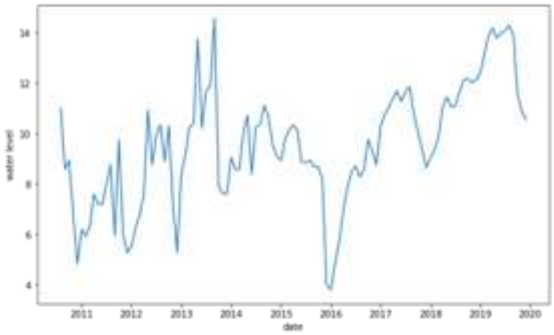
(P, D, Q): denotes the (p, d, q) for the seasonal part of time series.

V. Building Model

Data Visualization



(i)Chennai

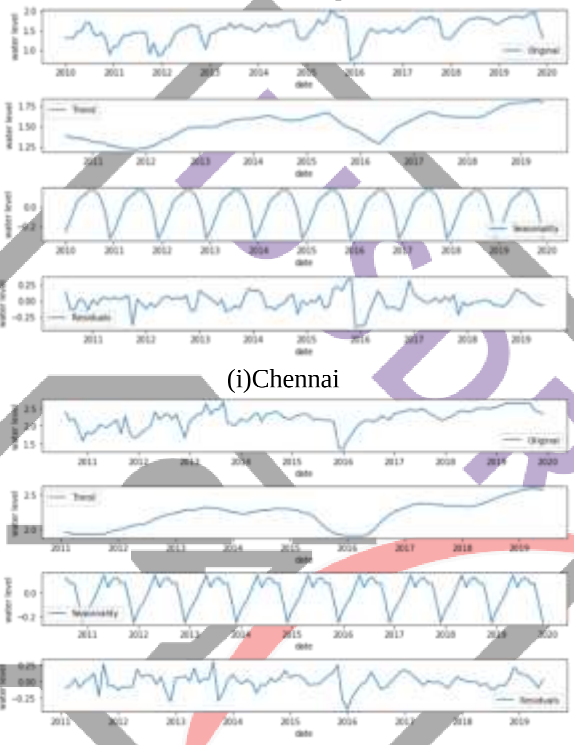


(ii)Dharmapuri

Figure 1: Ground Water Table (2010-2019)

Graphs in figure 1 shows the opted regions ground water level for the past decade, these are going to be used to develop the forecasting model.

The graphs in figure 2, decomposes the time series in its various components i.e. trend, seasonality, and residuals.



(i)Chennai

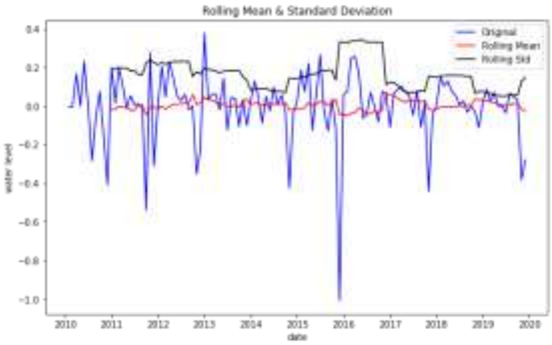
(ii)Dharmapuri

Figure 2: Decomposition of Time Series

From the graph in figure 2, we could clearly see that in both of the time series there is high seasonal behavior with a period of 12 months. Hence, instead of using ARIMA we could use SARIMAX which could provide better accurate results.

Identifying the Parameters

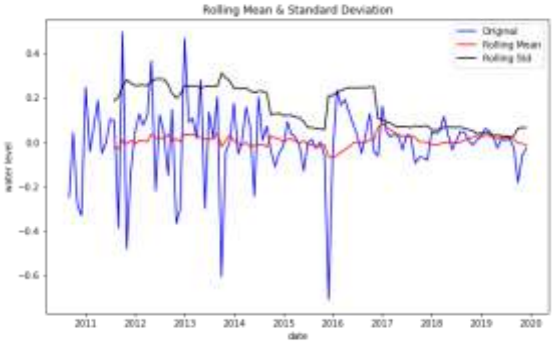
The first step of the ARIMA model building is to determine whether the variable is stationary across the time series. Stationary means that the values of the variable over time vary around a constant mean and variance. The graph in figure 2 clearly shows that the data or time series is not stationary. The series should be stationary in order to build an ARIMA model. First the time series have to be differenced 'd' times to obtain a stationary series in order to have an ARIMA (p, d, q) model with 'd' as the order of differencing. Precaution to be taken in differencing process as over differencing will tend to increase in the standard deviation, rather than a reduction.



Results of Dickey-Fuller Test:

```
Test Statistic                  -3.558307
p-value                          0.006610
#Lags Used                      12.000000
Number of Observations Used    106.000000
Critical Value (1%)             -3.493602
Critical Value (5%)             -2.889217
Critical Value (10%)            -2.581533
```

(i)Chennai



```
Results of Dickey-Fuller Test:
Test Statistic                  -2.969240
p-value                          0.037869
#Lags Used                      11.000000
Number of Observations Used    100.000000
Critical Value (1%)             -3.497501
Critical Value (5%)             -2.890906
Critical Value (10%)            -2.582435
```

(ii)Dharmapuri

Figure 3: Plot of 1st Order Differences and its Dickey-Fuller Test Results

It can easily be inferred from the graph in figure 3 that the time series appears to be stationary both in its mean and variance. But before proceeding further, the differenced time series data should be checked for stationary (unit root problem) using augmented Dickey-Fuller test.

Test for Stationarity: Augmented Dickey-Fuller (ADF) Test

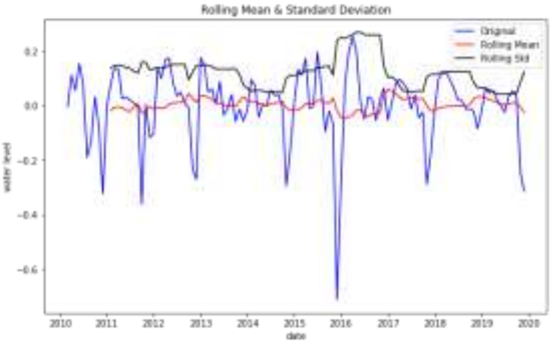
The null hypothesis (H_0) in the test is that the time series is not consistent i.e. non-stationary while the alternative hypothesis (Ha) is that the series is stationary. Then the hypothesis is tested by performing appropriate differencing of the data in d^{th} order and by applying ADF test to the differenced time series. A table of differenced data of current with the immediate previous one ($X_t = X_t - X_{t-1}$) is said to be the 1st order differentiation.

The ADF test result, as obtained are shown in Figure 3. As this fails to accept the H0, the alternative hypothesis is true i.e. the series is stationary in its mean and variance which infers to stop further differentiation of the time series. Hence, the value of ‘d’ is set to one (d = 1) for this ARIMA (p, d, q) model, but that not the case for Dharmapuri as it doesn’t satisfy the test. Since the comparative magnitude is very less and second order difference makes the time series more in its standard deviation, the value of ‘d’ is set to one (d=1) for Dharmapuri also.

Moving Average

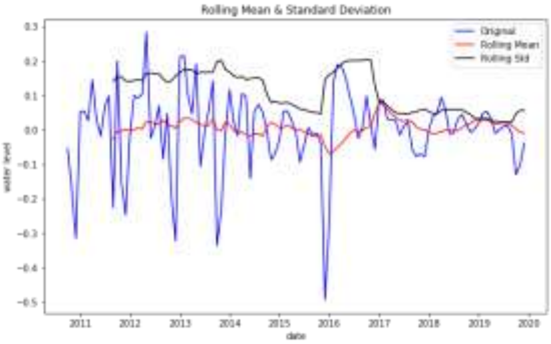
Past time point error terms are used to estimate the current and future point observations. Moving Average (MA) eliminates undetermined or random movements from the time series. Property 'q' represents the moving average in ARIMA. This is expressed as MA (x), where x represents the previous observations used to calculate the current.

After taking a moving average of three, the respective errors are found and a series is formed which is shown in figure 4.



```
          Results of Dickey-Fuller Test:
Test Statistic                  -2.887097
p-value                          0.046869
#Lags Used                      12.000000
Number of Observations Used    105.000000
Critical Value (1%)             -3.494220
Critical Value (5%)             -2.889485
Critical Value (10%)            -2.581676
```

(i)Chennai



```
          Results of Dickey-Fuller Test:
Test Statistic                  -2.874640
p-value                          0.048377
#Lags Used                      12.000000
Number of Observations Used     98.000000
Critical Value (1%)             -3.498910
Critical Value (5%)             -2.891516
Critical Value (10%)            -2.582760
```

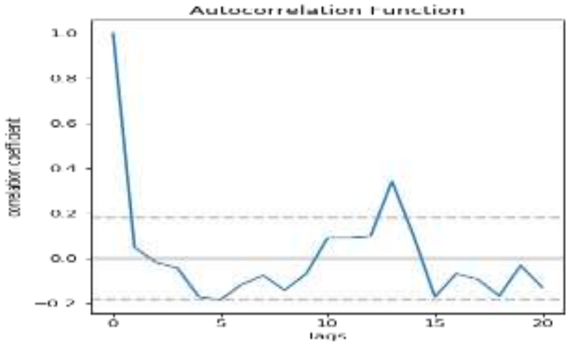
(ii)Dharmapuri

Figure 4: Moving Average's Dickey-Fuller Test

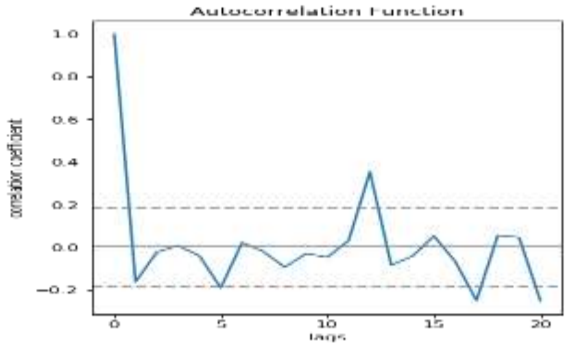
Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) Plots

Autocorrelation refers to how the time series correlates with its past values, while the ACF is the plot used to see the correlation between points, up to and including the lag unit. Typically, in the ARIMA model, either the AR or MA is used. In rare cases only both are used simultaneously. The ACF plot determines which of these terms to be used for time series analysis. The AR model is used if there is positive autocorrelation at log 1. If there is negative autocorrelation at log 1, then the MA model is used.

In this case the ACF plot of both regions have a positive autocorrelation at lag 1 that can be clearly seen in figure 5 and earlier the error series of the moving average of both regions also doesn't satisfied the dickey fuller test which could be seen in figure 4. Hence MA model is not considered and the value of 'q' is set to zero(q=0) for both the cases.

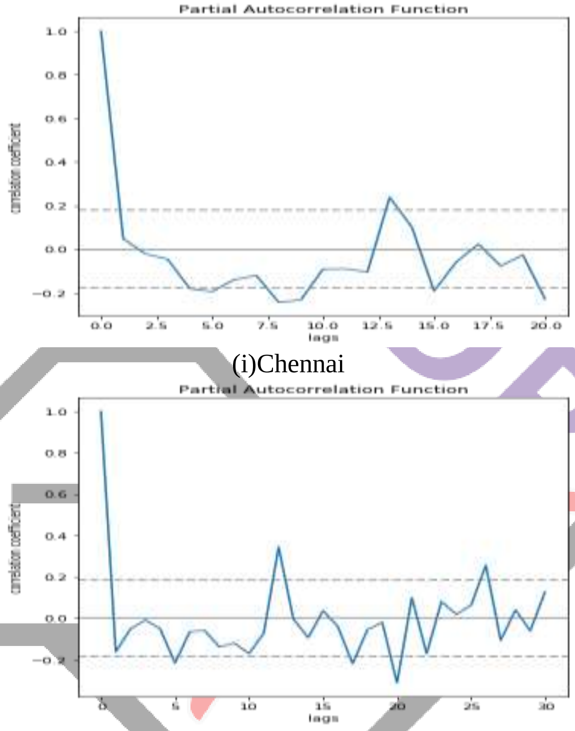


(i)Chennai



(ii)Dharmapuri
Figure 5: ACF Plots

To go further in steps for ARIMA model development, it is necessary to find suitable values of p for AR and for that partial autocorrelation function (PACF) should be examined. Partial autocorrelation is defined as the summary of the relationship between observations in the time series with observations at prior time steps associated with eliminated interventions. If the PACF plot falls on log n, the AR (n) model is used, and if the PACF decay is more gradual, the term MA is used.



(i)Chennai

(ii)Dharmapuri
Figure 6: PACF Plots

As in figure 6(i) the PACF drops off significantly at 1, p = 1 is taken as the value for Chennai region, but in case of Dharmapuri i.e. in figure 6(ii) it has more than one significant drops apart from 1 and since time series used here is of monthly wise and a high seasonality with a period of 12 months can be seen in figure 1, the value of ‘p’ could be probably taken as the next significant value i.e. p = 12. This also covers the seasonal effect to some extent.

ARIMA’s (p, d, q) parameters and SARIMAX’s (p, d, q) parameters are identical to each other. Only the seasonal order of SARIMAX i.e. (P, D, Q)m are remaining to be found.

It is obvious that the value of ‘m’ would be 12 because of 12 months per year and since MA model is already dropped, the value of ‘Q’ is set to 0. In both the model ‘D’ i.e. the seasonal order of differences would be more or less equal to ‘d’ i.e. the non -seasonal order of differences, as both would achieve stationarity after equal times of differences. In case of P, it is the seasonal autoregressive part of the time series. For Chennai region, the value of ‘P’ can be taken as one(P=1) as that would predict current month’s water level depending on the water level of previous year of the same month and for Dharmapuri region ‘P’ can be assigned to three(P=3) as its time series is much more deeply auto correlated, p = 12 i.e. the nonseasonal auto correlation of its series proves that.

Therefore, the final model developed are

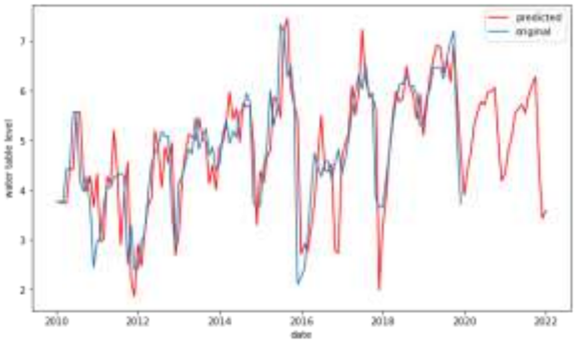
I ARIMA (1, 1, 0) (1, 1, 0)12 for Chennai
II ARIMA (12, 1, 0) (3, 1, 0)12 for Dharmapuri

VI. Predictions

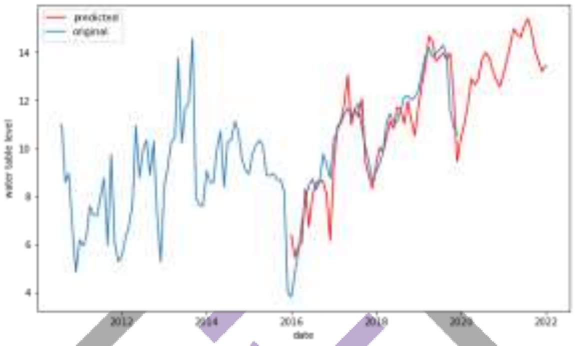
Using the SARIMAX model developed, water table level for the next two years of their respective regions are predicted in monthly basis. The table 1 shows the output delivered by the model. The water level is measured in meters below ground level i.e. mbgl.

Table 1: Forecasted Values

DATE	CHENNAI (mbgl)	DHARMAPURI (mbgl)
2020-01-01	3.89308631	10.41918125
2020-02-01	4.425196166	11.01930199
2020-03-01	4.764408267	11.77634043
2020-04-01	5.330401042	12.90924885
2020-05-01	5.585256666	12.64206781
2020-06-01	5.783847329	12.88748286
2020-07-01	5.710320501	13.69043535
2020-08-01	5.977830222	13.99002765
2020-09-01	5.984886352	13.79456983
2020-10-01	6.066468856	13.25835972
2020-11-01	5.046298896	12.85611561
2020-12-01	4.189333551	12.55811116
2021-01-01	4.324453562	13.02980518
2021-02-01	4.831221067	13.67169315
2021-03-01	5.056562452	14.25714288
2021-04-01	5.546730862	14.98032192
2021-05-01	5.64219235	14.75496128
2021-06-01	5.721268137	14.62337454
2021-07-01	5.553573378	15.10293117
2021-08-01	5.918872216	15.40220478
2021-09-01	6.114262719	14.91871963
2021-10-01	6.279146974	14.1164837
2021-11-01	4.494774014	13.63325922
2021-12-01	3.431365205	13.20548777
2022-01-01	3.58933391	13.44759792



(i)Chennai



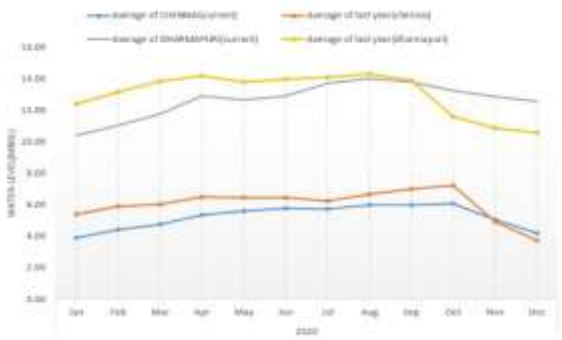
(ii)Dharmapuri

Figure 7: Original and Predicted Values

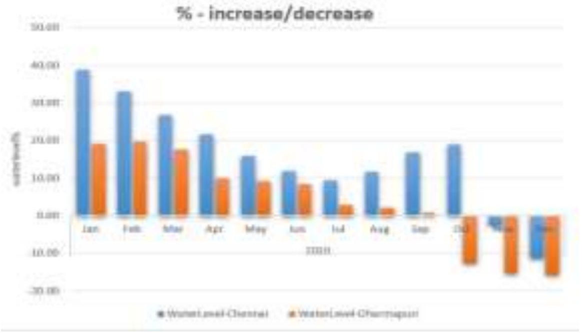
The developed model almost well mapped the previous year's water level which is illustrated in the figure 7. The red line represents the forecasted values and the blue line represents the input data values.

VII. Result and Discussion

As per the figure 8(ii), in Chennai the overall groundwater level is decreased by 15% compared to its previous year. The water availability also depends on the rate of consumptions. The rate of consumption is much higher which is obvious in case of Chennai (150 million liters per day). Though water level decreases, because of the much higher consumption rate scarcity and drought prevail later in the year.



(i)Chennai



(ii)Dharmapuri

Figure 8: Rise/Fall in Water Level

Coming to Dharmapuri the region was already in critical position as its water table level is much higher, about twice the Chennai's water table level. The recharge rate of the region is much lower which stands as the main reason for its continuous increase in water table level or drought.

VIII. Conclusion

The groundwater recharge rate of Chennai has not gone far worst compared to Dharmapuri. With artificial groundwater recharge, deficit in recharge rate can be controlled and that would help to achieve a stable water table level. Awareness and proper management of groundwater resources has to be engaged to decrease over utilization and wastage of water resources, especially groundwater resource as it is very minimal in quantity. Industries should be banned from consuming groundwater as they are the main cause of higher consumption rate. This would prevent occurrences of drought in Chennai.

In case of Dharmapuri region precipitation or rainfall are the only major measure to decrease the water table level as its natural recharge rate is very low. To increase precipitation forestation should be adopted in higher range.

References

[1] Debasish Sena and Naresh Kumar Nagwani, "A time-series forecasting-based prediction model to estimate groundwater levels in India".

[2] Salvatore Carta, Andrea Medda, Alessio Pili, Diego Reforgiato Recuperu and Roberto Daia, "Forecasting E-commerce products prices by combining an Autoregressive Intergrated Moving Average(ARIMA) model and Google trends data", 2018.

[3] Huanhuan Li, Yudong Lu, Ce Zheng, Mi Yang, Shuangli Li, "Groundwater level prediction for the arid oasis of northwest China based on the artificial bee colony algorithm and a back-propagation neural network with double hidden layers", 2019.

[4] Amir Jalalkamali, Hossein Sedghi and Mohammad Manshouri, "Monthly groundwater level prediction using ANN and neuro-fuzzy models: a case study on kerman plain, Iran".

[5] State ground and surface water resources data center.

[6] Groundwater Level Prediction/Forecasting and Assessment of Uncertainty Using SGS and ARIMA Models: A Case Study in the Bauru Aquifer System (Brazil).

[7] Jason Brownlee," Introduction to time series forecasting with python, prepare data and develop models to predict the future".